US-PAT-NO: 5809543

DOCUMENT-IDENTIFIER: US 5809543 A

TITLE: Fault tolerant extended processing complex for redundant nonvolatile file caching

DATE-ISSUED: September 15, 1998

INVENTOR-INFORMATION:

NAME	CITY	STATE	ZIP CODE	COUNTRY
Byers; Larry L.	Apple Valley	MN	N/A	N/A
Torgerson; James F.	Andover	MN	N/A	N/A
Price, deceased; Ferris T.	late of Mayer	MN	N/A	N/A

US-CL-CURRENT: 711/162, 711/120 , 711/167 , 714/14 , 714/6

ABSTRACT:

An outboard file cache extended processing complex for use with a host data processing system for providing closely coupled file caching capability is described. Data movers at the host provide the hardware interface to the outboard file cache, provide the formatting of file data and commands, and control the reading and writing of data from the extended processing complex. Host interface adapters receive file access commands sent from the data movers and provide cache access control. Directly coupled fiber optic links couple each of the data movers to the associated one of the host interface adapters and from the nonvolatile memory. A nonvolatile memory to store redundant copies of the cached file data is described. A system interface including bidirectional bus structures and index processors that control the routing of data signals, provides control of storage and retrieval of file cache data derived from host interface adapters and from the nonvolatile memory. Multiple power domains are described together with independent clock distribution within each power domain. The independent clock distribution sources are synchronized

with each other. A system for fault tolerant redundant storage of file cache data redundantly in at least two portions of the nonvolatile file cache storage is described.

45 Claims, 76 Drawing figures

Exemplary Claim Number: 15

Number of Drawing Sheets: 52

---------- KWIC ---------

Brief Summary Text - BSTX (6):
The relationship between the throughput rate of a data processing system, input/output (I/O) intensity, and data storage technology is discussed in "Storage hierarchies" by E. I. Cohen, et al., IBM Systems **Journal,** 28 No. 1 (1989). The concept of the storage hierarchy, as discussed in the article, is used here in the discussion of the prior art. In general terms, the storage hierarchy consists of data storage components within a data processing system, ranging from the cache of the central processing unit at the highest level of the hierarchy, to direct access storage devices at the lowest level of the hierarchy. I/O operations are required for access to data stored at the lowest level of the storage hierarchy.

Brief Summary Text - BSTX (14):
The third disadvantage associated with SSDs remains because two SSDs are required if fault tolerant capabilities are required. Fault tolerance with SSDs involves coupling two SSDs to a data processing system through two different data paths. A **backup** SSD mirrors the data on the primary SSD and is available in the event of failure of the primary SSD. To keep the **backup** SSD synchronized with the primary SSD, the instruction processor must perform two write operations when updating a file: the first write operation updates the primary SSD, and the second write operation updates the **backup** SSD. This method adds additional overhead to the data processing system to the detriment

of the system throughput rate.

Brief Summary Text - BSTX (44):

According to the present invention, the foregoing and other objects and advantages are attained by coupling an outboard file cache to a host file data processing system. The host issues file access commands which include a logical file-identifier and a logical offset. The outboard file cache includes a file descriptor table and cache memory for electronic **random access storage** of the cached files. The file descriptor table stores the logical file-identifiers and offsets of the portions of the files in the cache storage. Cache detection logic is interfaced with the file descriptor table and receives file access commands from the host. The file descriptor table is used to determine whether the portion of the file referenced by the file access command is present in the cache memory. Cache access control is responsive to the cache detection logic, and if the portion of the file referenced in the cache access command is present in cache memory, the desired access is provided. The outboard file cache is non-volatile relative to the main memory of the host because it is a separately powered storage system. Neither the host nor the outboard file cache is required to map the file data referenced in a file access command to the physical storage device and the physical address of the backing store on which the file data is stored if the referenced data is present in cache storage.

Detailed Description Text - DETX (18):

FIG. 4 illustrates an Outboard File Cache in a data storage hierarchy. A plurality of Control Units 104 labelled 104-I . . . 104-N, are coupled to Host 10 via IOPs 38-1 and 38-2 for providing access to Disks 106-1, 106-2, . . . 106-P and 106-N1, 106-N2, . . . 106-NQ. Application and system software executing on Host 10 reads data from and writes data to Files 108a-h. While Files 108a-h are depicted as blocks it should be understood that the data is not necessarily stored contiguously on the Disks 106. The Disks provide a backing store for retaining the Files. In the storage hierarchy, disks would fall into the category of secondary storage, with **primary storage** being the main memory of a Host.

Detailed Description Text - DETX (31):
The outboard file cache XPC 102 is configured with redundant power, redundant clocking, redundant storage, redundant storage access paths, and redundant processors for processing file access commands, all of which cooperate to provide a fault tolerant architecture for storing and manipulating file data. The outboard file cache XPC 102 is powered by dual Power Supplies 222a and 222b, which provide independent power domains within the XPC. The portion of the XPC to the left of dashed line 224 is powered by Power Supply 222a and is referred to as Power Domain A, and the portion of the XPC to the right of dashed line 224 is powered by Power Supply 222b and is referred to as Power Domain B. Each of Power Supplies 222a and 222b has a dedicated battery and generator **backup** (not shown) to protect against loss of the input power source.

Detailed Description Paragraph Table - DETL (2):
Word Bit Definition
0 0-3 These bits are reserved.
0 4-7
IXP.sub.-- # identifies the last IXP which updated this File Descriptor. This flag is useful for troubleshooting. 0 8-15 The PATH.sub.-- ID indicates the Host Interface Adapter 214 that is in the process of destaging, purging, or staging the Segment. 0 16-31 SEGMENT FLAGS are used to indicate various characteristics of the selected Segment 503 referenced by the File Descriptor 508. The flags include the following: SEGMENT.sub.-- WRITTEN is set when the Segment has been updated via a **write command** since the Segment was assigned. This flag is cleared when the Segment is destaged. TOTAL.sub.-- SEGMENT.sub.-- VALID is set when all blocks within a Segment are valid. A Segment is valid when each block in the Segment contains the most recent copy of the user's data. SEGMENT.sub.-- DISABLED identifies when a hardware

error was discovered for the associated Segment. SPECULATIVE/ORPHAN is a context sensitive flag. If the RESIDENT.sub.-- FILE flag is set, then this flag indicates whether the Segment is an orphan Segment. If the RESIDENT.sub.-- FILE flag is not set, this flag indicates whether the Segment was speculatively allocated. SEGMENT.sub.-- UNAVAILABLE is used to indicate whether the Segment referenced by the File Descriptor is eligible for cache replacement (reassignment). If this flag is set, then cache replacement algorithm does not consider the referenced Segment for reassignment. When this flag is set, the HASH.sub.-- LINK points to the next Segment available for cache replacement SEGMENT.sub.-- BUSY is used to indicate whether a read or write operation is in progress for the referenced Segment. The flag is set when a command is decoded, and remains set until the BLOCKS.sub.-- WRITTEN.sub.-- TEMPLATE has been updated. PURGE.sub.-- PENDING is used to indicate that a PURGE command found the referenced Segment had been updated, and is presently waiting for the Segment to be destaged before purging the Segment. DESTAGE.sub.-- PENDING is used to indicate that a DESTAGE command is in process. The flag is set when a DESTAGE command is decoded and cleared when the corresponding DESTAGE COMPLETE command is decoded. STAGE.sub.-- PENDING is used to indicate that a READ or **WRITE command** resulted in a miss condition, the Segment has been assigned, and the Segment is busy until the data has been written to the Segment. ALLOCATED.sub.-- WRITE.sub.-- MISS this flag indicates that the segment was assigned by either an ALLOCATE command or a **WRITE command.** SEQUENTIAL.sub.-- SEGMENT is set when multiple Segments are staged together or where the Segment immediately preceding the Segment is a Segment with the same FILE.sub.-- IDENTIFIER. The flag is used for determining which Segments should be destaged as a group. RESIDENT.sub.-- FILE indicates whether the Segment belongs to a Resident File. STICKING.sub.-- MASTER indicates whether the Host 10 has specified that the Segment should have a longer lifetime in the cache than Segments whose

STICKING.sub.-- MASTER flag is not set. NAIL is set when a Segment is not eligible for reassignment. The Index Processor 236 sets the NAIL flag for a segment for segments which are Nailed and segments which belong to Resident files. HOSTNAIL is set when a Segment in Nail Space has been created by the ALLOCATE command. PRE-USE is set by an IXP 236 to prevent another IXP from using the Segment. This flag indicates that an IXP has reserved the Segment so that the Segment is immediately available for assignment by the IXP. 1-2 FILE.sub.-- IDENTIFER identifies the File 106 to which the Segment is assigned. 3 FILE.sub.-- RELATIVE.sub.-- SEGMENT.sub.-- OFFSET indicates the location of the Segment relative to the first Segment in the file. 4 HASH.sub.-- LINK / BADPTR / NAIL.sub.-- LINK is the pointer to the next File Descriptor in a linked list of File Descriptors. If the SEGMENT.sub.-- UNAVAILABLE flag is set, the value in this field is used as the BADPTR, which is a pointer to the next Segment whose BAD.sub.-- OR.sub.-- UNAVAILABLE.sub.-- AREA is not set. If the NAIL flag is set, then the value in this field is used as the NAIL.sub.-- LINK which points to the next File Descriptor for a nailed Segment. 5 0-20 DATA.sub.-- POINTER is the physical address in NVS 220 where the Segment is stored. It is fixed at initialization and always points to the same segment. 5 21-27 FLAG ANNEX contains more flags which indicate characteristics of the Segment 503 referenced by the File Descriptor 508. The flags include the following: STICKING.sub.-- SLAVE is used to indicate the number of times the round robin cache replacement processing should exclude the referenced Segment from consideration for replacement. DESTAGE.sub.-- REPORTED is used to ensure that the IXP does not make more than one request for the Segment to be destaged. NEW is set if the Segment is within K Segments from selection for reassignment by the cache replacement algorithm. K is equai to one-half the number of Segments available in Cache File Space 522. NOTEPAD is a flag which has multiple uses. These uses will become apparent in the detailed discussion of the IXP processing. 5 28-31 BPID is the Back Panel Identifier associated with the NVS 220 in which the Segment is located. 6-7 BLOCKS.sub.-- WRITTEN.sub.-- TEMPLATE contains one bit for each block in the Segment. If a bit is set, it indicates that at some time after the Segment was last destaged, the corresponding block was updated. Bit 0 of Word 6 corresponds to Block 504-0 of a Segment 503, Bit 1 of Word 6 corresponds to Block 504-1 of Segment 503, . . . , Bit 31 of Word 6 corresponds to Block 504-31 of Segment 503, Bit 0 of Word 7 corresponds to Block 504-32 of Segment 503, . . . , and Bit 31 of Word 7 correspouds to Block 504-63 of Segment 503. 8 0-7 HOST.sub.-- ID is a value identifying the Host 10 that is in the process of destaging, purging, or staging the Segment. 8 8-15 GROUP.sub.-- ID indicates the group of Hosts 10 that are able to destage the Segment. In particular, the Group Identifier is the group of Hosts 10 that have direct access to the Disks 106 identified by the LEG1.sub.-- DISK.sub.-- NUMBER and LEG2.sub.-- DISK.sub.-- NUMBER. The group of Hosts 10 identified by the Group Identifier is called a "destage group." There are three types of destage groups: local, shared, and global. If the Group Identifier equals 0, then the Segment belongs to the global destage group; if the Group Identifier equals 1, then the Segment belongs to a local destage group; and if 2 &lt; = Group Identifier &lt; = 255, then the Segment belongs to a shared destage group. The number of local destage groups is equal to the number of Hosts 10 which are coupled to the outboard file cache XPC 102. There are 255 possible local destage groups. A Segment which is assigned to a local destage group can only be destaged by the Host 10 to which that local destage group is assigned. Note that if GROUP.sub.-- ID = 1, the HOST.sub.-- ID contained in the FILE.sub.-- IDENTIFIER must not equal zero and must specify a connected Host 10 that is able to destage the Segment. Otherwise, an error state has occurred. There are 254 possible shared destage groups. The set of Hosts 10 contained in a shared destage group is defined by the Host 10 software. The particular Hosts 10 contained in each shared destage group is dependent upon the Hosts 10 which are coupled to the outboard file cache XPC 102, the Disks 106 which are shared between the Hosts 10, and the particular files shared among the Hosts 10. 8 16-23 FILE.sub.-- SESSION is used for recovery

purposes when a Host fails unexpectedly. This field is beyond the scope of this invention. 8 24-31 HOST.sub.-- SESSION is Host Session Number in which the Segment was assigned to a file belonging to the Host. The Host Session Number is used for recovery purposes when a Host fails unexpectedly. This field is beyond the scope of this invention. 9 0-31 LEG1.sub.-- DISK.sub.-- NUMBER identifies the first disk on which the Segment is stored. "Leg" refers to the I/O Path on which the disk resides. 10 0-31 LEG2.sub.-- DISK.sub.-- NUMBER identifies the second disk on which the Segment is stored. 11 LEG1.sub.-- DISK.sub.-- ADDRESS specifies the address on the leg-1 disk at which the Segment is stored. 12 LEG2.sub.-- DISK.sub.-- ADDRESS specifies the address on the leg-2 disk at which the Segment is stored. 13-14 These words are unused.
15 PROGRAM.sub.-- ID identifies the Outboard File Cache program issued by a Host 10 that is in the process of destaging, purging, or staging the segment.